Poisoning Unauthorised Facial Recognition Models Using Image Cloaking

Abhishek Bhardwaj
Semester 5
Cluster Innovation Centre
University Of Delhi
abhishek02bhardwaj.er@gmail.com

Arpita Kesharwani
Semester 5
Cluster Innovation Centre
University Of Delhi
arpita151103@gmail.com

Siddhartha Mahajan
Semester 5
Cluster Innovation Centre
University Of Delhi
siddharthamahajan03@gmail.com

***Abstract*—This paper introduces a novel image cloaking approach utilizing MTCNN for facial recognition. The method involves the application of imperceptible perturbations to recognized faces, rendering them indistinguishable to human perception while confounding facial recognition models. Through rigorous experimentation, we demonstrate the effectiveness of this approach in protecting individual privacy, presenting a promising solution in the face of escalating facial recognition challenges.**

***Index Terms*—Facial Recognition, Privacy Preservation, Image Cloaking, MTCNN Framework, Deep Learning**

I. Introduction

In recent years, the pervasive integration of facial recognition systems has ushered in a new era of technological convenience but has also raised profound concerns regarding individual privacy. The capacity to effortlessly identify and track individuals through facial recognition algorithms poses significant risks, as highlighted by instances of unauthorized data harvesting, model training without explicit consent, and the potential for widespread surveillance. This surge, however, comes with profound privacy implications, exemplified by companies like Clearview.ai. Clearview.ai, through its controversial practices, demonstrated the capability to train powerful facial recognition models using publicly available images, raising ethical concerns about privacy infringement without individuals' knowledge or consent.

In response to these challenges, this paper introduces an innovative image cloaking technique that goes beyond traditional approaches. Our method employs the Multi-task Cascaded Convolutional Networks (MTCNN) for facial recognition, a widely utilized framework known for its accuracy in detecting and aligning faces in images. Building upon this foundation, our approach distinguishes itself by introducing imperceptible perturbations to the recognized faces.

The primary objective is to create cloaked images that remain visually unchanged to the human eye but significantly alter the features perceived by facial recognition models. This deliberate manipulation serves as a powerful countermeasure, enabling individuals to share images online without the risk of being accurately identified by automated recognition systems. To validate the efficacy of our approach, we conducted a series of comprehensive experiments, addressing three critical use cases. Firstly, we explore the scenario where a new Google Image account is populated exclusively with cloaked images and subsequently presented with an uncloaked image to assess if the facial recognition system recognizes them as distinct individuals. Secondly, we investigate the resilience of our cloaking method by creating an account filled with uncloaked images, probing whether the facial recognition model can differentiate between a cloaked and an uncloaked image. Finally, we delve into the adversarial potential of our approach by repeatedly injecting cloaked images into a repository dominated by uncloaked images, testing the capability of our method to introduce confusion and potentially poison the facial recognition model.

The ensuing sections of this paper provide a detailed exposition of our image cloaking methodology, the experimental framework, and the compelling results that underscore the practical viability of our proposed solution. As the relentless march of facial recognition technologies continues, our approach represents a timely and effective means for individuals to reclaim control over their online visual identity.

II. Background and Related Work

Facial recognition (FR) technology has become increasingly prevalent in recent years, offering convenience and security benefits in various applications. However, concerns about its potential to infringe upon individual privacy have spurred research into adversarial machine learning (AML) techniques that can protect identities from unauthorized recognition. This literature review examines existing AML approaches and their effectiveness in safeguarding privacy against FR systems.

A. *White-box vs. Black-box Attacks*

AML attacks can be categorized based on the adversary's knowledge of the target FR model:

White-box: The adversary has full access to the model's architecture and loss function, enabling manipulation of image pixels using techniques like Fast Gradient Sign Method (FGSM) and iterative methods [1].

Black-box: The adversary only observes the model's input-output behavior, necessitating more sophisticated techniques like transfer learning and evolutionary algorithms to generate effective adversarial examples .

B. *Privacy-Preserving AML for FR*

Several research works propose AML strategies to protect individuals from FR systems:

Fawkes: This system applies cloaks to images, transforming them into unrecognizable versions for FR models. However, its effectiveness relies on data poisoning attacks, making it vulnerable to model updates and detection mechanisms. [2]

Ulixes: This novel approach generates cloaks based on the target FR model's loss function, offering a more resilient defense against evolving FR technologies. Additionally, Ulixes boasts faster computation and less visual distortion compared to Fawkes. [3]

Other Related Works: FoggySight [4] and Face-Off [5] explore various attack methods and loss metrics, contributing valuable insights into the landscape of privacy-preserving AML for FR.

III. Methodology

A. *Multi-task Cascaded Convolutional Networks (MTCNN)*

Our image cloaking methodology leans on the robust architecture of Multi-task Cascaded Convolutional Networks (MTCNN), a specialized neural network designed for face detection and alignment. [6] MTCNN unfolds through three pivotal stages:

1) Stage 1 (P-Net): Proposals for potential face regions are generated by a fully convolutional network (FCN). These proposals undergo refinement via bounding box regression and non-maximum suppression.

2) Stage 2 (R-Net): Further scrutiny is applied to the refined proposals by a more complex network, enhancing face localization accuracy.

3) Stage 3 (O-Net): The final stage refines and validates face regions, providing precise facial landmarks and bounding box coordinates. MTCNN operates adeptly by scanning images at multiple scales, demonstrating exceptional accuracy in detecting and aligning faces of various sizes and orientations.

B. *Image Cloaking Methodology*

1) ***Assumptions and Threat Model Recap:*** Our image cloaking methodology is devised to address the privacy concerns of users sharing images online, with a specific focus on thwarting third-party trackers and unauthorized facial recognition models. The defense mechanism is intricately tailored to fortify against potential privacy infringements.

2) ***Cloaking Objectives:*** The core objectives of our image cloaking methodology revolve around the introduction of imperceptible perturbations, denoted as "cloaks," to user images. Formally, let x represent a user image, and x_T denote a target image. The primary goal is to create visually indistinguishable images to humans while inducing significant alterations in features perceived by facial recognition models.

3) ***Technical Aspects of Cloaking:***

a) ***Cloak Perturbations:*** For each user image x, the corresponding perturbation (cloak) is computed based on a target image x_T. Mathematically, the optimization objective can be formulated as follows:

$$\text{argmax}_\delta \ \text{Deviation}(x, x_T + \delta)$$

subject to the constraint $||\delta||_p \leq \rho$, where Deviation represents the feature representation deviation, and ρ is the perceptual perturbation budget.

b) ***Efficient Search with Landmark:*** Another image x_T from a distinct user class serves as a landmark for an efficient search during the optimization process. This ensures that the cloak is guided by a dissimilar target, enhancing its effectiveness against recognition models.

c) ***Image-Specific Cloaking:*** Cloaks are generated in an image-specific manner, guaranteeing diversity among cloak patterns for user images. The optimization minimizes the Euclidean distance between the feature representation of the target image x_T and the cloaked image $x + \delta$.

d) ***Cloaking Effectiveness & Transferability:*** Even in the absence of the target class (T), cloaking induces misclassification due to dissimilarity in feature representations. The transferability property ensures the methodology's efficacy against different architectures and training data.

C. *Optimisation Problems*

In the realm of privacy-preserving image manipulation, the optimization challenge arises when crafting an image $(x_A + \delta)$ that appears authentic to human observers as person A, yet strategically misleads face detection models into recognizing it as person B. The optimization problem is formulated as:

$$\min_\delta \text{Perceptual\_Diff} + \lambda \times \text{Model\_Mismatch}$$

Here, Perceptual_Diff measures the perceptual dissimilarity between the original (x_A) and perturbed $(x_A + \delta)$ images, Model_Mismatch quantifies the dissimilarity between the perturbed image's features and those of person B, and λ is a balancing parameter.

Armed with moderate computing resources, users can locally apply cloaking using the MTCNN framework for facial recognition. The detailed implementation encapsulates the technical aspects, offering a systematic procedure for image cloaking. Real-world limitations and privacy considerations are acknowledged, providing a comprehensive exploration of the experimental framework, empirical results, and the practical viability of the proposed image cloaking solution.

IV. Results and Discussion

To rigorously evaluate the susceptibility of facial recognition systems to poisoned data, we conducted comprehensive experiments using Google Photos, a popular platform leveraging facial recognition features for image categorization.

A. *Case 1: No Prior Database*

We initiated the test by exclusively populating Google Photos with 50 deliberately manipulated, poisoned images, devoid of any existing database. Our objective was to train the system solely on poisoned data, excluding any genuine images. Upon introducing authentic, unpoisoned images into the system, Google Photos demonstrated a striking inability to recognize these unaltered images. The system's failure rate stood at 100% across 50 recognition attempts for genuine images, highlighting a severe vulnerability to poisoned data during training.

B. *Case 2: Malicious Entity with a Prior Database*

In this scenario, we simulated a situation where the adversarial entity introduced a combination of 30 unpoisoned and 20 poisoned images into the Google Photos database, mimicking a pre-existing dataset comprising both authentic and manipulated data. Subsequently, during recognition accuracy assessments, Google Photos displayed a discernible decline in accuracy correlated with the increased presence of poisoned images. Notably, the accuracy rate proportionally decreased as the quantity of poisoned images augmented. The reduction in accuracy signified the system's susceptibility to degraded performance when contaminated with poisoned data, indicating a notable vulnerability.

C. *Case 3: Poisoned Image Recognition with a Prior Database*

To further explore the system's robustness, we trained Google Photos with 50 unaltered, authentic images, simulating a genuinely constituted database. Following the training phase, we introduced a single poisoned image for recognition. Intriguingly, the system consistently failed to identify the poisoned image, exhibiting a 92% failure rate across recognition attempts. This result underscores a significant inability of the system to detect manipulated data within an authentic dataset.

D. *Overall Implications*

The conducted experiments vividly demonstrate the susceptibility of facial recognition systems, exemplified by Google Photos, to the detrimental effects of poisoned data. Whether introduced during training or recognition phases, the presence of manipulated data severely impairs the system's ability to accurately identify unaltered, genuine images. These findings underscore the urgent need for enhanced defenses and robust countermeasures to mitigate the adverse impact of adversarial attacks on facial recognition technology. Securing these systems against poisoned data is imperative for bolstering their accuracy, reliability, and trustworthiness in real-world applications.

Conclusion

In conclusion, our proposed image cloaking methodology, leveraging the robust Multi-task Cascaded Convolutional Networks (MTCNN) framework, presents a novel and effective approach to counter the growing challenges posed by facial recognition systems. By introducing imperceptible perturbations to recognized faces, our method aims to provide individuals with a potent tool to protect their privacy in an increasingly interconnected digital landscape.

The comprehensive experiments conducted on Google Photos highlight the susceptibility of facial recognition systems to the influence of poisoned data. Whether introduced during training or recognition phases, the presence of manipulated data severely impairs the system's accuracy in identifying unaltered, genuine images. These findings underscore the urgent need for enhanced defenses and robust countermeasures to mitigate the adverse impact of adversarial attacks on facial recognition technology.

As we navigate the complexities of facial recognition technologies, our approach stands as a timely and effective means for individuals to reclaim control over their online visual identity. However, it's important to acknowledge the real-world limitations and privacy considerations associated with any privacy-preserving technique.

Future research endeavors should explore the scalability and adaptability of our approach to diverse facial recognition models and real-world scenarios. Additionally, addressing the ethical implications and potential countermeasures against adversarial attacks remains an essential avenue for further exploration.

In conclusion, our image cloaking methodology offers a promising solution in fortifying individual privacy against unauthorized facial recognition models. As the technological landscape continues to evolve, proactive measures to safeguard individual privacy become increasingly crucial.

Acknowledgment

We extend our sincere gratitude to Dr. Anjani Verma for invaluable mentorship, providing technical expertise, and steadfast support throughout our project. Dr. Verma's guidance has been instrumental in shaping our research trajectory, enriching our academic experience.

References

[1] Lin, J., Xu, L., Liu, Y., & Zhang, X. (2020). Black-box adversarial sample generation based on differential evolution. In Journal of Systems and Software (Vol. 170, p. 110767). Elsevier BV. https://doi.org/10.1016/j.jss.2020.110767

[2] Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020, August). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. 29th USENIX Security Symposium (USENIX Security 20), 1589–1604. Retrieved from https://www.usenix.org/conference/usenixsecurity20/presentation/shan

[3] Cilloni, T., Wang, W., Walter, C., & Fleming, C. (2021). Ulixes: Facial Recognition Privacy with Adversarial Machine Learning. In Proceedings on Privacy Enhancing Technologies (Vol. 2022, Issue 1, pp. 148–165). Privacy Enhancing Technologies Symposium Advisory Board. https://doi.org/10.2478/popets-2022-0008

[4] Evtimov, I., Sturmfels, P., & Kohno, T. (2020). FoggySight: A Scheme for Facial Lookup Privacy (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2012.08588

[5] Chandrasekaran, V., Gao, C., Tang, B., Fawaz, K., Jha, S., Banerjee, S. (2021). Face-Off: Adversarial Face Obfuscation. In Proceedings on Privacy Enhancing Technologies (Vol. 2021, Issue 2, pp. 369–390). Privacy Enhancing Technologies Symposium Advisory Board. https://doi.org/10.2478/popets-2021-0032

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.